SDOC et TermWatch : deux méthodes complémentaires de cartographie de thèmes

Fidelia Ibekwe-Sanjuan, Xavier Polanco, Eric Sanjuan

SDOC et TermWatch : deux méthodes complémentaires de cartographie de thèmes

IBEKWE-SANJUAN Fidelia
ERSICOM
Université Jean – Moulin, Lyon 3
ibekwe@univ-lyon3.fr

POLANCO Xavier
Unité de recherche et innovation
INIST – CNRS
polanco@inist.fr

SANJUAN Eric
IUT STID – LITA (EA 3097)
Université de Metz
eric.sanjuan@univ-metz.fr

1. Introduction

L'objectif de cette communication est de comparer deux méthodes SDOC et TermWatch de classification hiérarchique non supervisée (*clustering*) de données textuelles, initialement destinées à la veille scientifique et technique dans une application de fouille de textes. Les deux logiciels proposent à l'utilisateur de visualiser les résultats sous forme d'une carte thématique. Elles sont cependant complémentaires :

- SDOC est fondé sur l'analyse de la matrice de co-occurence et positionne les classes (clusters) sur le plan en fonction de leurs propriétés structurelles. SDOC permet aussi de visualiser le contenu des classes sous forme de graphes. SDOC est un outil développé par l'Unité de Recherche et Innovation de l'INIST – CNRS, aujourd'hui accessible via le site WWW de STANALYST® (cf. Annexe jointe).
- TermWatch classifie directement une liste de termes, automatiquement extraits des documents, en fonction de leurs seuls liens de variation syntaxique et donc sans utiliser de notion d'occurrence d'unités textuelles dans les documents, ni de ressources terminologiques extérieures. Les résultats sont présentés sous forme d'un réseau que l'on peut visualiser et explorer avec le logiciel AiSee® (*http://www.aisee.com*). Les liens sont d'autant plus resserrés que les classes partagent de nombreuses variantes de termes et qu'elles sont donc supposées être thématiquement proches. TermWatch est une plateforme développée en collaboration par ERSICOM de l'Université LYON III et le LITA de l'Université de Metz.

Le rôle de la classification non supervisée étant d'assurer la tâche de dépistage et d'agencement des éléments thématiques au sein des données textuelles, les classes dégagées par ces deux méthodes permettent de réaliser une analyse thématique de l'information traitée. Un thème correspondant à une classe (cluster ou agrégat). Autrement dit, les classes sont censées représenter (signifier, désigner) des thèmes (ou des thématiques). Chaque classe est assimilable à un thème désigné par un libellé automatiquement extrait, ainsi que par les mots-clefs ou les termes constituant la classe. En outre, pour chacun de ces thèmes il est possible d'extraire les documents ou les textes de la base qui leur sont le plus associés et de faire correspondre les acteurs humains (auteurs) et institutionnels (laboratoires, entreprises, pays) qui sont responsables de leur développement.

L'expérimentation a été menée sur des données extraites à l'aide de STANALYST® de la base de données PASCAL maintenue par l'INIST-CNRS, pour la période 1997- 2003, en fonction de 16 revues scientifiques toutes de langue anglaise et réputées pour la publication d'articles en Recherche d'Information (Information Retrieval – IR). Il est cependant important de noter que le corpus de notices obtenu avec titres et résumés d'auteurs ne constitue pas un corpus de spécialité en IR puisque les revues dont elles sont issues couvrent une large

problématique autour de l'informatique et des sciences de l'information. On a obtenu ainsi 3 355 données bibliographiques (articles scientifiques).

Le tableau suivant présente les noms des revues sélectionnées précédés du nombre de notices extraites ainsi que de leurs contributions respectives à la constitution de ce corpus.

Tab. 1 : *Rang, nombre de notices, pourcentage, effectifs cumulés, pourcentages cumulés et nom des revues à l'origine de la constitution du corpus*.

1	831	25%	831	25%	Information sciences
2	688	21%	1519	45%	J. of the Am. Soc. for Information Science and Technology
3	283	8%	1802	54%	Information processing & management
4	272	8%	2074	62%	Journal of information science
5	267	8%	2341	70%	Information systems management
6	175	5%	2516	75%	Journal of Documentation
7	176	5%	2692	80%	Information Systems
8	116	3%	2808	84%	Information systems security
9	108	3%	2916	87%	Library & information science research
10	108	3%	3024	90%	Online information review
11	87	3%	3111	93%	Journal of internet cataloging
12	70	2%	3181	95%	Information retrieval & library automation
13	67	2%	3248	97%	Knowledge organization
14	44	1%	3292	98%	Journal of Information Science and Engineering
15	34	1%	3326	99%	International forum on information and documentation
16	29	1%	3355	100%	Information retrieval
	3355	100%			

Il est important de remarquer que les deux seules revues de rang 1 et 2 sont à l'origine de presque la moitié des notices extraites tandis que la revue spécialisée en IR *Information retrieval* est celle qui contribue le moins au corpus.

A chaque publication (texte) est associé une liste de mots clés en signalant le contenu (les connaissances contenues dans le texte en question). Cette indexation est fournie avec les données textuelles et elle représente un total de 5 199 mots-clés dont 3 186 mots-clés de fréquence égale à 1, c'est-à-dire 61,28 % de ce vocabulaire d'indexation. La moyenne de mots-clés par document est de 9,13. Cette indexation a été utilisée pour l'analyse de mots associés utilisant le programme SDOC du module INFOMETRIE de STANALYST.

Par contre cette indexation n'est pas utilisée par TermWatch. Ce sont es champs textuels, titres et résumés d'auteurs, qui ont été directement utilisés pour une extraction de termes (syntagmes nominaux) avec INTEX. L'ensemble de ces textes à une taille totale de *454 412* mots et couvre un vocabulaire de *20 928* mots distincts. Usuellement la liste des termes extraits est préalablement soumise pour filtrage et validation par un expert avant d'être transmise à TermWatch. Dans le cadre de cette expérimentation en fouille de texte, nous avons choisi de soumettre au système l'intégralité des candidats termes extraits au nombre de 50 737, sans aucun filtrage préalable, ni humain, ni statistique.

Voici le plan du reste de la communication. Nous commençons par rappeler sommairement le principe des deux méthodes, SDOC en section 2 et TermWatch en section 3. Pour plus de détails sur ces méthodes nous renvoyons le lecteur aux références données en fin d'artivcle. En section 3 nous utilisons SDOC pour obtenir une carte thématique exhaustive du corpus présenté ci-dessus. En section 4 nous illustrons comment TermWatch permet de faire émerger le sous-réseau thématique à l'origine du choix des revues, à savoir le thème transversal d' l'IR. Enfin, en section 5, nous montrons que TermWatch a permis de dégager une thématique

émergeante en IR, la théorie des Rough Sets, que nous retrouvons dans la classification générée par SDOC.

2. SDOC et la méthode de mots associés

Le programme SDOC est une implémentation de la méthode de mots associées. La méthode de mots associés repose sur la comptabilisation des co-occurrences de mots indexant les différents documents d'un corpus. L'idée est qu'un domaine peut être identifié par son propre vocabulaire ou terminologie, ou plus exactement par les associations qu'une spécialité établit entre des mots. Plus de mots co-occurrent fréquemment dans des textes différents et plus les connaissances indexées par ces paires de mots et les connexions qu'ils représentent se renforcent. Les informations concernant la fréquence d'occurrence et de co-occurrence des mots peuvent être synthétisées dans une matrice carrée où les lignes et les colonnes désignent des mots-clés. Sur la diagonale de la matrice, se trouve la fréquence de chaque mot-clé dans le corpus ; dans les autres cases, le nombre de co-occurrences de chaque paire de mots-clés. La matrice symétrique obtenue est appelée matrice de distances : m_i est le nombre d'occurrences de chaque mot *i* dans l'ensemble d'articles scientifiques, m_j est le nombre d'occurrences de chaque mot *j* dans l'ensemble et m_{ij} le nombre de co-occurrences de chaque paire de mots.

Réseau des associations

Pour construire les réseau des associations de mots, la première étape consiste à calculer le nombre d'occurrences m_i de chaque mot *i* dans l'ensemble d'articles et le nombre de co-occurrences m_{ij} de chaque paire de mots m_i et m_j. Cependant, la co-occurrence ne permet pas à elle seule de mesurer la force des associations entre les mots, car elle avantage les mots apparaissant un grand nombre de fois par rapport aux autres. On calcule donc un coefficient d'association normalisé :

$$A_{ij} = m_{ij}^2 / m_i * m_j$$

c'est-à-dire la co-occurrence au carré des mots *i* et *j*, divisée par le produit de leurs fréquences respectives. Il vaut 0 si les mots *i* et *j* n'apparaissent jamais simultanément et 1 dans le cas inverse. Dans ce cas, on a l'égalité : $m_{ij} = m_i = m_j$. Par ailleurs, ce coefficient est analogue aux indices bien connus de Dice, de Jaccard et de Salton.

Classification thématique

L'étape suivante est de procéder à une classification thématique des données textuelles. Dans le but de dégager les connaissances enfouies dans l'ensemble de données, une classification automatique non supervisée est utilisée. SDOC applique un algorithme de classification ascendante hiérarchique (CAH) du simple lien, afin de construire des classes de mots associés proches les uns des autres n'excédant pas une taille maximale. Le paramétrage de la taille donne lieu au fait qu'il y ait à la fois des associations internes aux classes (intra-classes) et des associations externes ou inter-classes. Ainsi, cette méthode permet d'identifier à la fois les classes et les liens qui les unissent. Après le processus de classification des mots-clés, les documents sont affectés aux classes.

Densité et centralité des classes

Chaque classe est caractérisé par sa *densité* intra-classe et sa *centralité* dans le réseaux de classes (inter-classe). Les notions de densité et de centralité sont destinées à mettre en évidence la contribution des différentes classes à la structuration du réseau général.

La *densité* d'une classe est la moyenne de ses associations internes, c'est un indicateur (il indique, signale, désigne) la cohérence de la classe au sens où elle mesure la force des liens qui unissent les mots composants de la classe. La densité vise à caractériser l'intensité des

liens qui unissent les mots qui composent une classe. Les classes peuvent être rangées par ordre de densité décroissante (ou au contraire par ordre de densité croissante).

La *centralité* d'une classe est la moyenne de ses associations externes ; elle indique le degré d'association de cette classe aux autres. La centralité rend compte pour une classe de l'intensité de ses liens avec d'autres classes. Le mesure de la centralité permet de ranger les différents classes (agrégats, clusters) par ordre de centralité décroissante (ou au contraire par ordre de centralité croissante).

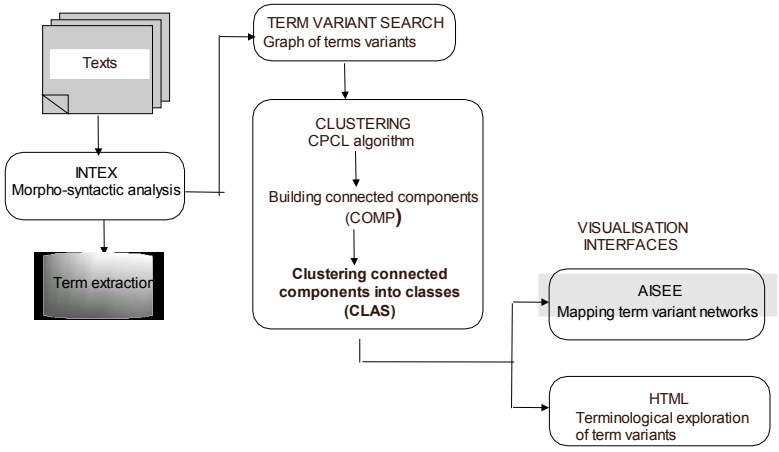
Représentation cartographique des classes

Les valeurs de densité et de centralité des classes permettent de construire une représentation cartographique. Les classes sont ainsi représentées dans un espace bi-dimensionnel défini par ces deux indicateurs : la centralité (X) et la densité (Y). La carte est obtenue en rangeant les classes (clusters) horizontalement suivant l'axe X par ordre de centralité croissante, et verticalement suivant l'axe Y par ordre de densité croissante. Cette opération permet de classer, à l'aide des valeurs médianes, tous les classes en quatre catégories qui correspondent aux quatre quadrants de la carte. Le fait que deux classes soient proches l'une de l'autre sur la carte ne signifie pas qu'elles soient liées l'une à l'autre, ceci montre seulement que leurs indices de centralité et de densité ont des valeurs voisines. En d'autres termes, la proximité entre deux clusters indique qu'ils sont structurellement proches, mais ne présage pas de leur proximité sémantique.

3. TermWatch et la réduction de graphes syntaxiques

TermWatch est aussi un système de classification non supervisée de données textuelles, sauf qu'il ne classe pas des mots-clefs indexant des documents, mais des termes. Il s'agit actuellement de syntagmes nominaux directement extraits avec le logiciel INTEX de textes en langue anglaise (ici les titres et les résumés d'auteur), comprenant au moins un centre (sujet) et un modifieur. Initialement destiné à la VST (Veille Scientifique et Technique), ce système a été jusqu'ici appliqué à des ensemble de textes scientifiques et techniques ayant un certaine homogénéité thématique. Autre différence avec SDOC, ce système présente la carte de thématiques qu'il produit sous forme d'un graphe que des logiciels spécialisés récents tels que AiSee peuvent disposer sur le plan en fonction de leurs seules propriétés structurelles (cliques, cycles, sous-arborescences, etc.). La figure suivante rappelle l'architecture du système.

Fig. 1 *: la plateforme TermWatch*



Réseau de termes

Une série d'automates et de transducteurs a été conçu dans INTEX pour extraire les termes. Ces automates sont successivement et automatiquement appelées par un script shell. On récupère une liste de syntagmes nominaux de plusieurs mots susceptibles de désigner de par leurs propriétés syntaxiques et grammaticales, un objet ou une notion du domaine (donc un terme). Ils sont spécifiques au domaine considéré et il n'est pas rare qu'ils n'aient qu'une unique occurrence dans tout le corpus de textes.

La notion d'association entre termes, dans un but de classification non supervisée, est entièrement fondée sur les relations de variation syntaxiques que peuvent partager ces termes. Cela représente une intéressante alternative à l'utilisation de la co-occurrence comme critère d'association, tout particulièrement lorsqu'il s'agit d'extraire une information rare. Les termes sont ainsi insérés dans un réseau de variantes. Le tableau suivant donne les relations de variation extraites sur ce corpus et utilisées pour la classification.

Tab. 2 : *Nom, description et effectif (nombre de couples de termes en relation) des relations de variation extraites.*

nom	*description*	*effectif*
expansion gauche	ajout de mots avant le premier des modifieurs	*8 892*
insertions	insertion de modifieurs à une unique position du terme	*4 327*
expansion droite	ajout de mots après le centre (sujet) du terme	*7 859*
expansion double	application simultanée de l'expansion gauche et droite	*5 217*
substitution de centre	substitution du centre dans les termes de longueur 3	*15 037*

Le système extrait en fait bien d'autres relations de variation, c'est l'utilisateur qui choisit celles qu'il veut utiliser pour classification. Ici, comme nous avons décidé de travailler sur une liste non filtrée de candidats termes, nous nous sommes restreints aux relations supposés être les moins bruyantes, c'est à dire avec moins de *100 000* liens sur l'ensemble du corpus.

Classification non supervisée

La méthode utilisée dans *TermWatch* dissocie les types de variations en deux catégories selon qu'ils provoquent ou pas un changement de centre (sujet) :

- COMP qui comprend ici expansion gauche et insertion,
- CLAS qui comprend ici Expansion droite, expansion gauche-droite et substitution de centre.

On utilise la première catégorie COMP pour former des composantes connexes de termes, au sens de la théorie des graphe et la deuxième catégorie CLAS pour lier ces composantes et former des thématiques. Ces composantes représente le premier niveau de classification. Leur taille peut varier de 2 à 1 000 et leur regroupement nécessite la mise en oeuvre d'un algorithme spécifique qui ne néglige pas les plus petites composantes connexes.

Pour cela il utilise un indice de dissimilarité *d* qui à tout couple de composantes connexes, associe la somme des proportions de liens de variation des relations dans CLAS, entre ces deux composantes connexes. Plus formellement, *d* est une application dans [0,1] définie pour tout couple (*i, j*) de composantes connexes de la manière suivante :

i) $d(i,j) = 1$ si pour tout r dans $\{1,\ldots,k\}$, $N_r(i,j) = 0$;

ii) $d(i,j) = 0$ si i = j ;

iii) $d(i,j) = 1 / \sum_{r=1}^{k} \frac{N_r(i,j)}{|R_r|}$ où $R_1 \ldots R_k$ désignent les relations dans CLAS et $N_r(i,j)$ est le nombre de liens dans R_r entre *i* and *j*.

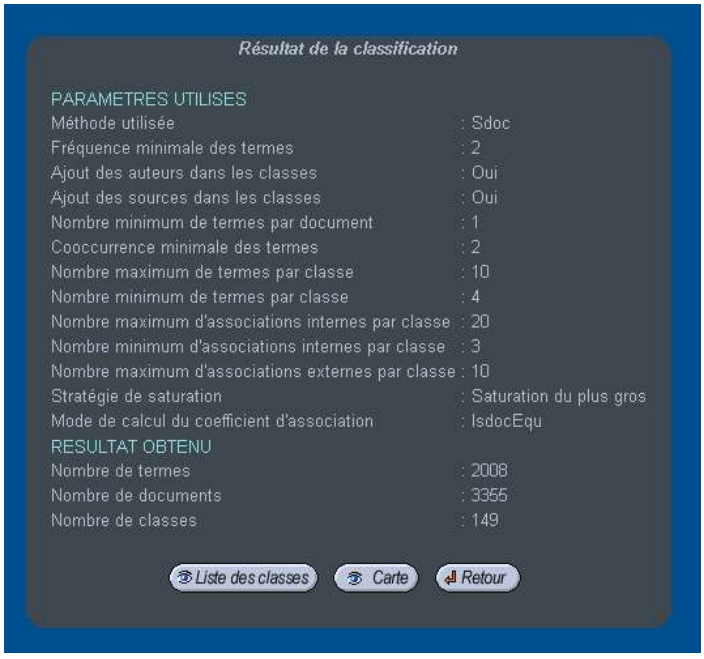
Utiliser cette dissimilarité *d* pour agglomérer les composantes multi-termes en classes par classification ascendante hiérarchique revient à approcher *d* par une ultramétrique *u*, de choisir un niveau significatif du dendrogramme et de visualiser le graphe obtenu en agglomérant les composantes dans une même classe. Il est bien connu que la meilleure approximation inférieure serait l'ultramétrique associée à la classification par lien simple (CLS). Mais l'objectif n'étant pas la meilleure approximation numérique de *d* mais celle qui préserve la structure du réseau de composantes et éviter l'effet de chaîne propre à la CLS. TermWatch utilise alors un critère d'agglomération local qui consiste à agglomérer deux classes seulement si la dissimilarité entre elles est plus faible qu'avec toute autre classe dans leur voisinage.

Cette ultramétrique particulière a la propriété de dégager des classifications avec un nombre de classes non triviales (non réduites à un singleton) bien plus important que la CLS, tout en partagent la majorité de ses propriétés mathématiques, dont l'unicité.

4. Cartes des thématiques générées par SDOC

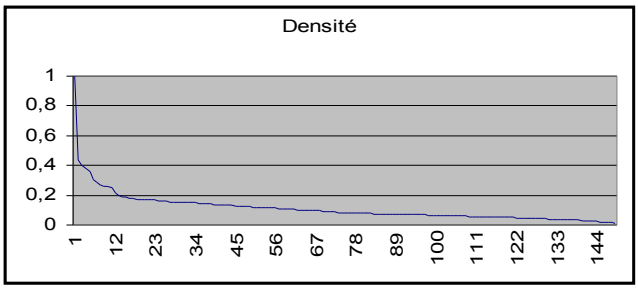
Nous présentons ici la classification générale extraite par SDOC. Sur le corpus de 3 355 documents indexés par 2 008 mots-clés de fréquence >= 2, SDOC a généré 149 classes. La figure suivante reproduit les paramètres de la classification.

Fig. 2 : *paramètrage employé*



Ce paramétrage permet d'obtenir une classification de granularité fine, c'est-à-dire une nombre significatif de classes. Dans la figure suivante les classes sont rangées par ordre de densité décroissante. C'est un courbe J dont le max = 1 et le min = 0.013.

Fig 3 : *Distribution des 149 classes par ordre de densité décroissante*



On voit sur cette figure que la majorité des classes se trouvent au-dessous du seuil *0,2* c'est-à-dire qu'elles présentent une faible densité. Comme nous l'avons déjà dit, la *densité* d'une classe est la moyenne de ses associations internes. La densité vise à caractériser l'intensité des liens qui unissent les mots qui composent une classe. Ainsi, la densité est considérée comme un indicateur de la cohérence de la classe au sens où elle mesure la force des liens qui unissent les mots composants de la classe.

On étudie de même la courbe de *Centralité*, calculée sur les classes rangées par ordre de centralité décroissante. C'est également un courbe J avec *0,318* comme maximum et *0* comme minilum. Cette courbe J montre que la plus grande partie des classes (au total *139*) ont un faible indice de centralité (inférieur à *0,100*). Comme nous l'avons aussi déjà dit, la *centralité* d'une classe est la moyenne de ses associations externes ; elle indique le degré d'association de cette classe aux autres. En d'autres termes, la centralité rend compte pour une classe de l'intensité de ses liens avec d'autres classes.

Chaque classe étant définissable par sa centralité et par sa densité, il est donc possible de tracer une carte. La carte est obtenue en rangeant les classes (clusters) horizontalement suivant l'axe X par ordre de centralité croissante, et verticalement suivant l'axe Y par ordre de densité croissante. Cette opération permet de classer, à l'aide des valeurs médianes, tous les classes en quatre catégories qui correspondent aux quatre quadrants de la carte. Les classes de type 1 ont à la fois des valeurs de densité et de centralité élevées. Ces classes constituent en quelque sorte le centre du domaine. Les classes de type 2 ont une centralité forte, mais la densité de leurs liens internes est relativement faible. Les classes de type 3 présentent une forte densité mais une faible centralité. Les classes de type 4 sont à la fois de centralité et de densité faibles.

On obtient de cette manière la carte reproduite en figure 4. La répartition des classes (thèmes) dans le quadrant 4 indique que le domaine analysé s'organise autour d'un réseau de thèmes qui sont à la fois de densité et de centralité faibles. En somme, qu'il s'agit d'un domaine faiblement structuré. Ce qui peut s'expliquer par la manière dont le corpus a été constitué. Le thème de la recherche d'information, à l'origine du choix des revues, apparaît bien mais non central, ni dense. Il est noyé dans la multiplicité des thèmes abordées par les différentes revues.

La figure 5 suivante montre la même carte, mais cette fois ci paramétrée par rang, ce qui produit un effet de zoom sur les classes de la carte, notamment des classes groupées dans le quadrant 4. Ceci permet de visualiser surtout la distribution des classes de type 4 à nouveau en en quatre catégories. Cette fois-ci on retrouve bien la classe Information Retrieval dans le premier cadran.

Fig. 4 : *La Carte globale*

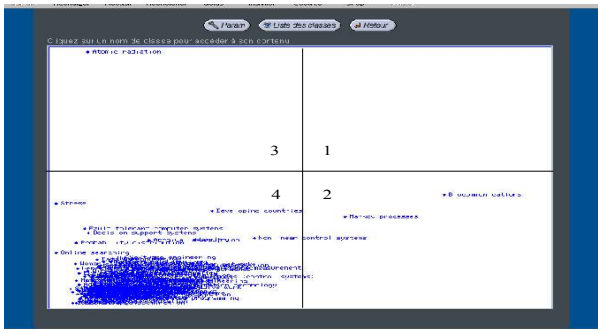
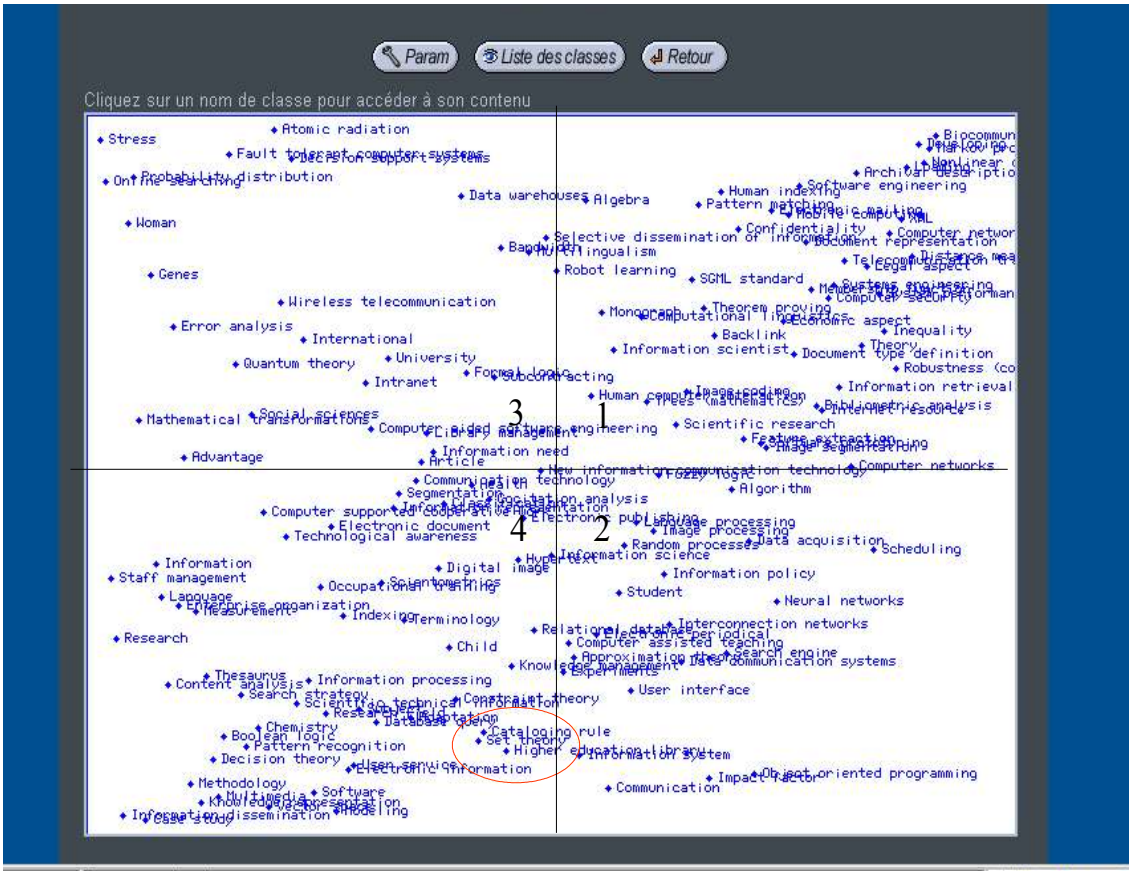


Fig. 5 : *La Carte globale par rang (effet de zoom)*

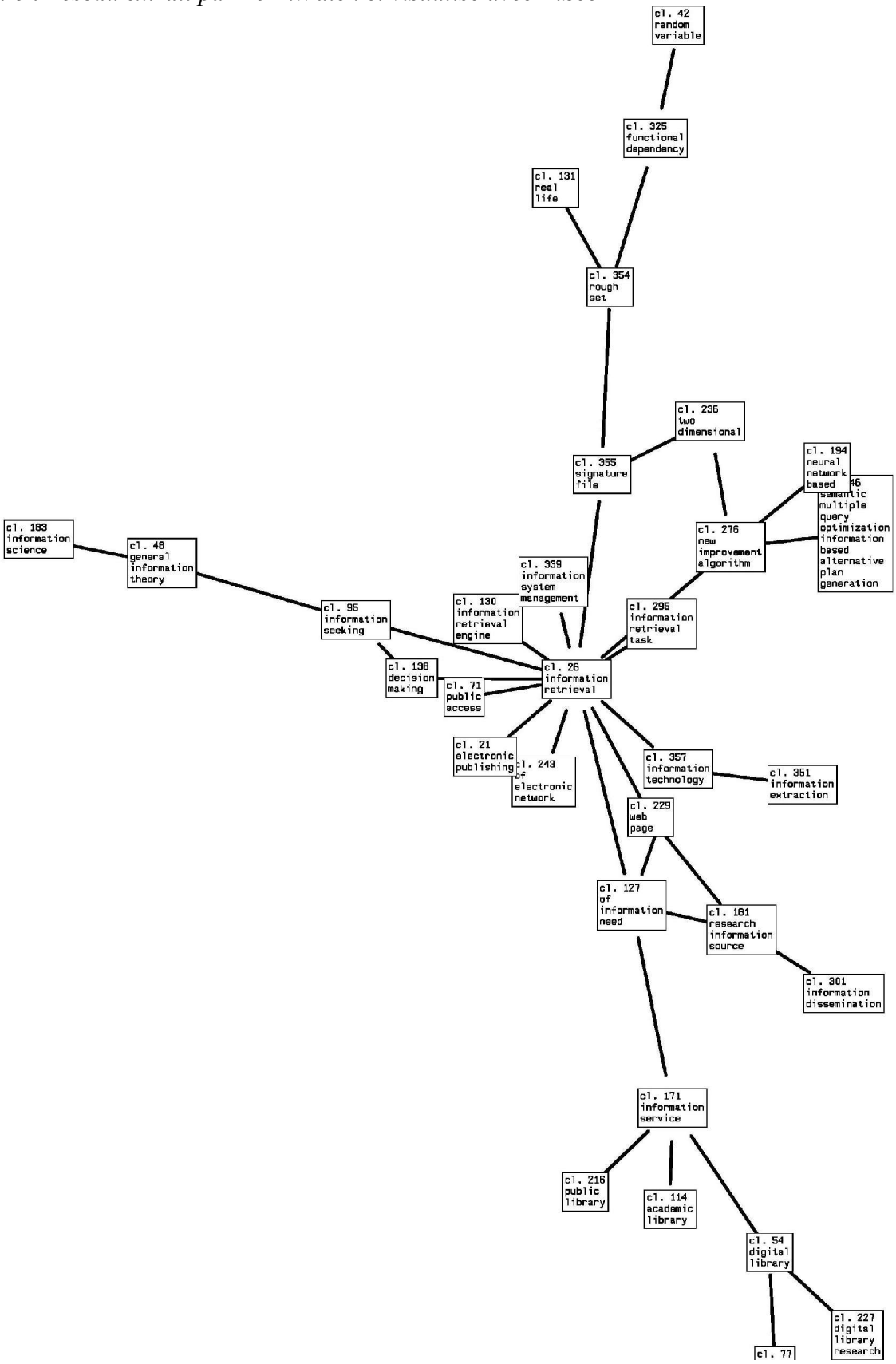


Sur cette carte la classe Set Theory est signalée par un cercle rouge et est commentée plus loin.

5. Extraction d'un sous réseau transversal avec TermWarch

L'algorithme de classification de TermWatch décrit ci-dessus a été itéré deux fois (deuxième niveau de la CAH). A cette étape *6 849* termes regroupés en *4 252* composantes connexes ont été classées en *397* classes. Les classes sont automatiquement libellées par le terme ayant la plus forte activité de variation. On a ainsi obtenu une classification de taille comparable avec celle obtenue par SDOC.

Fig. 6 : *Réseau extrait par TermWatch et visualisé avec AiSee*



Cependant lorsque le graphe obtenu est trop dense, l'utilisateur peut choisir de ne visualiser que les classes liées par des liens au dessus de certains seuils pré-calculés par le programme.

Nous avons ainsi choisi d'ignorer les liens en dessous du premier seuil qui correspond au double du lien le plus faible entre deux classes. La figure 6 précédente montre le graphe ainsi obtenu et visualisé avec AiSee. Il s'agit en fait d'un sous-graphe de la classification globale. Sa forme étoilée centrée sur la classe libellée *information retrieval* est remarquable.

Le tableau suivant donne le nombre de termes (nb_term) pour chaque classe de la figure 6 avec le nombre de documents ayant au moins un terme dans la classe (nb_doc) et la proportion maximale de termes de la classe pouvant être contenu dans un même document (max_score). Le tableau est classé par nombre de documents.

Tab. 3 : *Libellés, effectifs, nombre de documents et couverture des classes de la Figure 6.*

Libellé	***nb_term***	***nb_doc***	***max_score***
information retrieval	997	1196	0.01
research information source	89	531	0.08
rough set	163	232	0.04
new improvement algorithm	192	221	0.04
signature file	202	220	0.04
information technology	71	207	0.1
need to find	110	198	0.05
information science	39	187	0.1
digital library	74	181	0.16
information seeking	100	153	0.08
web page	174	143	0.03
of information need	72	115	0.08
academic library	32	59	0.13
decision making	23	56	0.26
information retrieval task	56	56	0.09
digital library research	24	54	0.33
public library	23	52	0.3
information system management	32	49	0.19
of electronic network	26	44	0.15
real life	5	36	0.4
random variable	34	33	0.15
general information theory	28	29	0.25
information service	46	29	0.09
two dimensional	8	28	0.5
functional dependency	27	28	0.33
public access	7	27	0.43
neural network based	10	22	0.5
electronic publishing	7	15	0.29
semantic multiple query optimization information based alternative plan generation	13	13	0.46
information retrieval engine	5	11	1
information dissemination	16	11	0.19
information extraction	6	8	0.33

Lors des expérimentations précédentes de TermWatch sur des corpus constitués par des requêtes de Veille Scientifique et Technique, de taille réduite et filtrées, le contenu des classes était directement interprétable et lisible tandis que les libellés semblaient arbitraires puisque plusieurs termes d'une même classe pouvaient avoir une activité de variation maximale. Dans le cas présent nous sommes confrontés à la situation inverse. La majorité de ces classes sont de taille trop grande pour être directement lisibles. Par contre le critère de choix du libellé par activité de variation maximale devient déterministe et est pleinement justifié. Le nombre de termes reste cependant un bon indicateur de l'importance de la classe tandis que le nombre de documents ayant au moins un terme dans la classe renseigne sur la transversalité de la classe dans le corpus. Enfin le score maximal renseigne de la concentration de ces classes dans les documents.

On remarque alors que de nombreux libellées extraits par TermWatch recoupent des libellés de la carte visualisée avec SDOC. Ainsi on retrouve une classe libellée *Information retrieval,* le libellé *neural network based* dans les résultats de TermWatch renvoie au libellé *Neural NetWorks* de la carte produite par SDOC, *Random variable* renvoie à *Random processes, New improvement algorithm à Algorithm*, *decision making à décision theory* etc. Plus précisément, il est possible de relier la majorité des libellés extraits par TermWatch à des libellés de classes extraites par SDOC, en utilisant les mêmes relations de variation que celles détectées et utilisées par TermWatch. Cela n'était pas a priori évident puisque les libellés extraits par TermWatch proviennent d'une liste de termes contenant énormément de bruit compte tenu du parti pris de ne procéder à aucun filtrage pour cette expérimentation. Le recoupement de ces libellés permet alors une lecture simultanée des deux cartes.

D'autre part, le tableau précédent positionne la classe *information retrieval* comme étant la plus importante et la plus transversale. TermWatch a ainsi automatiquement fait émergé le sous-réseau de thématiques présentes dans ces revues, fédérées autour de la recherche d'information (IR) qui a motivé le choix de ces 16 revues.

On vérifie que ce sous réseau est sémantiquement cohérent. En effet, on retrouve la thématique IR cernée par les classes *information retrieval task*, *information retrieval engine* qui corespondent à un léger glissement de sujet : *moteur* et *tâche*. D'autres sujets tels que *electronic publishing*, *information technology*, *public access*, *decision making*, *information seeking* sont directement liés à la classe *Information Retrieval*. *Information extraction* un domaine plus récent de spécialisation de la recherche n'est pas directement relié au noyau du graphe mais à la classe *information technology*, ce qui pourrait refléter l'importance de l'informatique en ce domaine. Il est aussi intéressant de noter le positionnement de domaines reliés à la recherche d'information. Ainsi *information need* relie le sous-réseau formé par *information service* au noyau. Aussi ce dernier thème est à sont tour lié aux domaines *public library, digital library, academic library, digital library research.* La classe qui a été mal libellée *need to find* contient en fait les variantes de termes sur les ressources de sites WWW. Le libellé erroné est dû à une mauvaise analyse morphologique de “need”. La classe libellée *new improvement algorithm* et positionnée en haut du tableau précédent, contient quatre composantes libellées *dynamic programming, image processing, database application and new improvement algorithm* respectivement. La dernière des composantes contient des termes variants tels que *genetic programming algorithm, filtering algorithm, clustering algorithm, new hybrid algorithm* et décrit les développement de nouveaux algorithmes pouvant être appliqués à l'IR. Cette classe est logiquement proche de *neural network* qui contient des termes tel que *genetic algorithm technique, neural network based clustering technique.*

Le domaine de recherche historiquement le plus ancien en sciences de l'information se retrouve à l'une des extrémités du graphe dans la classe libellée *general information theory*

qui regroupe les composantes *general information theory, information science theory growth, theory growth tool analysis*.

6. Détection d'une thématique émergeante : la classe Rough Set

Plus surprenant et sans-doute plus intéressant d'un point de vue fouille de données est la révélation de la classe Rough Set classée en haut du tableau précédent et liée à la classe IR dans la figure 6. Pour comprendre ce positionnement il est nécessaire de retourner au contenu des textes ce que nous avons fait en immergeant les données de classification générées par TermWatch dans une Base de données relationnelle. Le tableau suivant donne les titres des sept documents ayant le plus de termes dans la classe Rough Set.

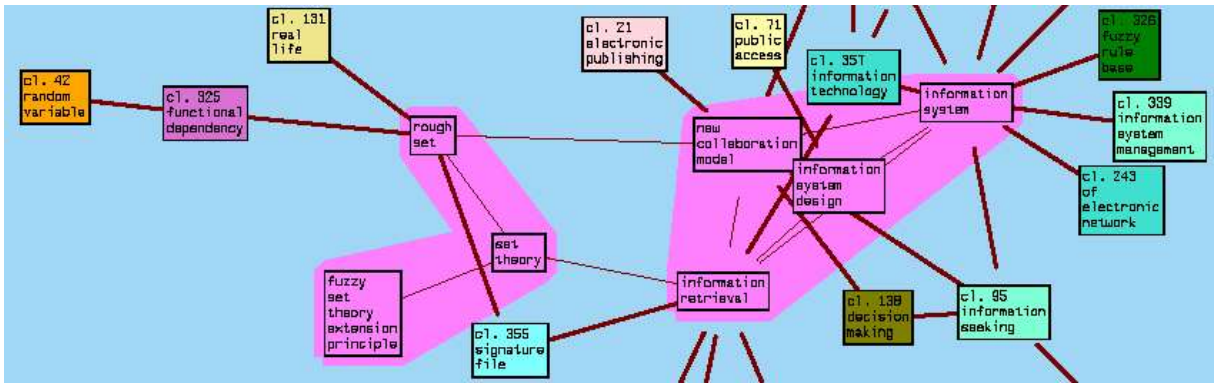
Tab. 4 : *Titres des documents les plus fortement associé à la classe Rough Set*

Rang	***Titre***	***Année***	***Nombre***
1	Validation of authentic reasoning expert systems	1999	7
2	Double-faced rough sets and rough communication	2002	6
3	Canonical forms of fuzzy truthoods by meta-theory based upon modal logic	2001	6
4	On axiomatic characterizations of crisp approximation operators	2000	6
5	alpha -RST: a generalization of rough set theory	2000	5
6	Application of rough sets to information retrieval	1998	5
7	Parallel fuzzy inference based on level sets and generalized means	1997	5

Premier point rassurant, sur ces sept documents, cinq traitent directement des Rough Sets. Il s'agit des documents 1, 2, 4,5 et 6. Cela n'était pas a priori évident du fait que la classe contient 232 termes et qu'elle a été constituée sans aucune information statistique sur la présence absence des termes dans les documents.

Les graphes générés par TermWatch sont codés de manière à ce que chaque sommet qui représente une classe puisse être éclaté en composantes de manière interactive avec l'interface AiSee. Celles-ci sont alors drapéee sur un même fond de couleur et positionnées en fonction des autres classes auxquelles elles sont reliées. La figure suivante montre le développement des classes Rough Set et Information Retrieval détectées par TermWatch.

Fig. 7 : *développement des classes Rough Set et Information retrieval*



On remarque ainsi que la classe Rough Set est constituée de thèmes liés à la théorie des ensembles et de ses extensions : Rough Set, Fuzzy Set and Set Theory.

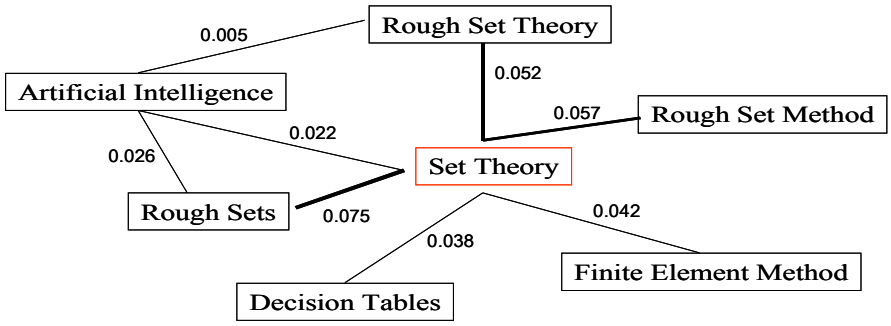
La mise en avant de cette classe par TermWatch est une bonne illustration du type de thématique émergeante que ce système peut traquer. En effet, la théorie des rough sets, initialement conçue pour des applications telles que les télécommunications et les systèmes experts, comme l'illustre le titre des deux documents les plus fortement associé à cette classe, a récemment été appliqué à l'IR comme le signale le titre du sixième document. C'est d'ailleurs le terme *Rough Set Theory to Information Retrieval* extrait de ce document qui est à l'origine du lien dans la figure 6 entre la classe Rough Set et IR.

D'autre part cette théorie admet désormais une extension “Fuzzy” introduite sous le nom de *alpha-RS*, nom qui est présent dans le titre du cinquième document à l'origine des variantes qui ont conduit à l'immersion dans une même classe des thèmes “Fuzzy” et “Rough”.

Cette classe émergeante ayant été détectée, on peut tenter de la retrouvée parmi les classes extraites par SDOC en utilisant les libellées des composantes la constituant. On retrouve alors bien le thème Rough Set représenté dans la classe Set Theory par *21* articles sur un total de *118* groupés dans cette classe. Cette classe agrège *229* auteurs et *118* articles scientifiques provenant de *6* périodiques.

Le contenu en mots clefs et les associations internes d'une classe sont synthétisées par Sdoc sous la forme d'un graphe reproduit ci-dessus.

Fig. 8 : *Graphe de la classe Set Theory dégagée par Sdoc*



On retrouve ainsi l'association entre les termes Rough Set et Set Theory signalée par TermWatch. Par contre le thème des Fuzzy Sets est occulté par la thème plus général d'Intelligence Artificielle.

D'autre part le tableau qui suit montre que cette classe extraite par Sdoc est fortement liée à la classe Theory par trois liens de poids 0.055, 0.036 et 0.027 et de co-occurrence 53, 55 et 25 respectivement, comme nous pouvons le voir dans la liste des associations externes de la classe. Ensuite, elle a des liens importants avec la classe Algebra.

SDOC a ainsi dégagé les thèmes majeurs des publications sur les Rough Sets et bien connus des spécialistes de cette théorie, à savoir des travaux algébriques et théoriques dont beaucoup sur les algèbres modales et multivaluées étudiées de manière intensive en intelligence artificielle. Par contre TermWacht a détecté ce vers quoi les concepteurs de cette théorie voudraient tendre, à savoir des applications à la « vie réelle » telles que l'IR. Cet objectif influe directement sur la terminologie qu'ils emploient ce qui explique le positionnement différent de cette même classe par TermWatch.

Tab. 3 : *Contenu de la Classe Set theory dégagée par Sdoc où se trouve le thème Rough Set*

Liste des termes composants de la classe :		
Poids (*)	Fréquence	Libellé
0.667	53	Set Theory
0.389	87	Artificial Intelligence
0.111	9	Rough Set Theory
0.111	4	Rough Sets
0.056	2	Decision Tables
0.056	4	Finite Element Method
0.056	3	Rough Set Method

Liste des associations internes :			
A_{ij}	Cooccurrence	Terme i	Terme j
0.075	4	Rough Sets	Set Theory
0.057	3	Rough Set Method	Set Theory
0.052	5	Rough Set Theory	Set Theory
0.042	3	Finite Element Method	Set Theory
0.038	2	Decision Tables	Set Theory
0.026	3	Artificial Intelligence	Rough Sets
0.022	10	Artificial Intelligence	Set Theory
0.005	2	Artificial Intelligence	Rough Set Theory

Liste des associations externes :				
Classe externe	A_{ij}	Cooccurrence	Terme i	Terme j
Theory	0.055	53	Set Theory	Theory
Image Processing	0.038	2	Set Theory	Rule Induction Methods
Algebra	0.038	2	Set Theory	Relational Grammar
Algebra	0.038	2	Set Theory	Natural Languages
Theory	0.036	55	Artificial Intelligence	Theory
Theorem Proving	0.028	11	Set Theory	Theorem Proving
Fuzzy Logic	0.028	7	Artificial Intelligence	Expert System
Theory	0.027	25	Artificial Intelligence	Application
Robustness (control systems)	0.026	3	Artificial Intelligence	Computational Verb System
Mathematical Transform.	0.025	2	Set Theory	Turing Machines

7. Conclusion

Ce retour d'expérience illustre l'intérêt pour la fouille de textes d'utiliser la complémentarité des approches fondées sur l'analyse de la co-occurrence et de graphes de termes. Sur ce corpus d'étude, SDOC a permis une extraction exhaustive des thématiques présentes dans les notices et les a situées en termes de centralité et densité. TermWatch a par contre permis de dégager de manière automatique et sans ressources extérieures le sous-réseau de thématiques transversales à l'origine du choix des revues qui ont constitué le corpus et au moins une thématique émergeante en la classe Rough Set, qu'il positionne en fonction de sa terminologie. Le caractère émergent de cette thématique a pu être vérifié par un retour au contenu des textes. Son étude a pu être poursuivie grâce à la découverte d'une classe extraite par SDOC portant sur cette même thème.

Références

Dowdall J., Rinaldi F., Ibekwe-SanJuan F., SanJuan E. *Complex structuring of term variants for Question Answering. Workshop on Multiword expressions : Analysis, Acquisition and Treatment.* In 41st Meeting of the Association for Computational Linguistics (ACL, 2003), Sapporo, Japan, 12 Juillet, 2003, 8p.

Ibekwe-SanJuan F., SanJuan E. *From term variants to research topics.* International Journal on Knowledge Organization (ISKO), Special issue on Human Language Technology, vol. 29, n° 3/4, 22p, 2003.

Ibekwe-SanJuan F., SanJuan E. *Cartographie de réseaux de termes.* 5ème Journées Terminologie et Intelligence Artificielle (TIA'03), Strasbourg, 31 mars -1 avril 2003, 124-134.

Ibekwe-SanJuan F., Dubois C., *Can Syntactic variations highlight semantic links between domain topics ?* 6th International Conference on Terminology and Knowledge engineering (TKE'02), Nancy 28-30 août 2002, p. 57-63.

Grivel L., Mutschke P., Polanco X., *Thematic Mapping on Bibliographic Databases by Cluster Analysis: A Description of the SDOC Environment with SOLIS*, Knowledge Organization, vol. 22, num. 2, pp. 70-77, 1995.

Jacquemin C., *Spotting and discovering terms through Natural Language Processing*, MIT Press, 378p, 2001.

Kodratoff Y. *Knowledge discovery in texts* : *A definition and applications*, in Foundation of Intelligent systems, Ras & Skowron (eds.) Lecture Notes in Artificial Intelligence, n° 1609, Springer, pp. 16-29, 1999.

Xavier P., François C., Royauté Jj., Besagni D., Roche I., *STANALYST: An Integrated Environment for Clustering and Mapping Analysis on Science and Technology* in Proceedings of the 8th International Conference on Scientometrics and Informatrics, Sydney, Australia, July 16-20th 2001, Vol. 2, pp. 871-873

Polanco X., Grivel L., Royauté J. *How to do things with terms in informetrics : terminological variation and stabilization as science watch indicators.* Proceedings of the 5th International Conference of the International Society for Scientometrics and Informetrics, Illinois USA, 7-10 June 1995, pp.435-444.

Sander G. *Visualisierungstechniken für den Compilerbau.* Dissertation, Pirrot Verlag & Druck, 1996.

Silberztein M. *Dictionnaire électronique et analyse automatique des textes.* Le système INTEX. Masson, Paris, 1993.

Annexe

STANALYST

La station d'analyse STANALYST® se compose d'un ensemble de modules applicatifs permettant la recherche d'informations dans des bases de données documentaires de l'INIST en vue de leur analyse statistique, terminologique et thématique. L'intégration de ces différents modules au sein d'une interface graphique commune, accessible depuis un navigateur http est la suivante : l'ACCUEIL est une page HTML statique qui permet l'accès à l'application, l'utilisateur déclare son nom et définit son mot de passe. Le PROJET permet de définir un environnement de travail, c'est-à-dire un répertoire dans lequel seront stockés tous les résultats concernant le projet. L'utilisateur est le propriétaire et il a également la possibilité de donner accès à son projet aux utilisateurs associés. Les modules CORPUS, BIBLIOMÉTRIE, INDEXATION et INFOMÉTRIE constituent les modules de travail de cette station d'analyse de l'information : Le module CORPUS gère la création de corpus par exécution de requêtes construites par l'utilisateur. Les corpus peuvent ensuite être exportés à destination des modules suivants. Le module BIBLIOMÉTRIE gère la création d'analyses statistiques descriptives. Le module INDEXATION permet de réviser l'indexation en vue de la classification thématique, ou bien de réaliser une indexation automatique du corpus, il s'appuie pour cela sur les outils ILC (permettant une extraction terminologique à partir de plusieurs référentiels terminologiques). Le module INFOMÉTRIE gère la classification à partir des outils de classification automatique non supervisée SDOC, NEURODOC. Les modules utilisent un ensemble de répertoires de travail contenant les programmes, scripts, paramétrages nécessaires à son fonctionnement, ainsi que l'ensemble des projets créés par les utilisateurs. Le schéma suivant résume l'architecture de la plate-forme :